Optimization of Value of Aerodrome Forecasts

ROSS KEITH*

Bureau of Meteorology, and School of Mathematical and Physical Sciences, James Cook University, Townsville, Australia

(Manuscript received 14 March 2002, in final form 29 January 2003)

ABSTRACT

Prediction of short-term variations of vital boundary layer conditions at airports, such as visibility and cloud base, is important to the safe and economic operation of airlines. Results of an experiment involving groups of forecasters at three different locations across Australia are presented. The forecasters were asked to indicate their degree of confidence that weather at the airport would be below ''minimums'' that would require aircraft to carry adequate fuel to proceed to an alternate destination should they not be able to land. The results of the trial are shown to closely obey a Gaussian model as used in signal detection theory (SDT). The data are fitted to an accuracy-value model developed by Mason. The paper demonstrates the ability of forecasters to provide reasonably reliable probability forecasts of significant events at airports. The potential value in reliable estimation of the probability of low visibility and cloud base at aerodromes is estimated by using cost parameters for two actual examples of flights into Melbourne and Townsville, Australia.

1. Introduction

The value of a weather forecast to a user is about much more than accuracy. The most important aspect of the design of a forecast delivery system is to optimize the flow of the opinion of the forecaster directly to the user. The user should then utilize this opinion in an optimal way, given the relative costs of occurrence of the event and protection from the event. In this paper it is assumed that the user, in this case an airline, would want to use the forecast in a way that optimizes the economic outcome without compromising safety.

This issue of obtaining the maximum value from the expertise of the forecaster is the main topic of this study. Meteorological authorities have gone to great lengths to measure weather forecast accuracy, but less focus has been given to forecast value. Many papers have been written on the subject of accuracy and value in weather forecasts (e.g., Murphy 1977, 1985; Katz and Murphy 1997). These describe the extra value inherent in probabilistic forecasts. More recently studies have been published on the value of ensemble numerical forecasting and its relative value with respect to increasing the resolution of models. Wilks (2001), Zhu et. al. (2002), and Richardson (2000) use broadly similar methodology in ascribing the value of probabilistic forecasts based on, inter alia, numerical ensembles, with respect to climatology and perfect forecasts. Of interest for the current study, Wilks found that forecasts exhibiting consistent overforecasting (i.e., a low decision threshold) produce greater value for situations with low cost–loss (C–L) ratios. This is shown to be the case in this study for airline costs. Most short-duration flights have very low C–L ratios. Wilks also found that the use of probabilistic methods achieved the greatest increase in value over climatology, with respect to perfect forecasts, for forecast probability densities typical of short-range forecasts. Zhu et al. use a relative operating characteristic (ROC) area summary measure to demonstrate the value of the multiple decision thresholds inherent in probabilistic ensemble methods.

However, little or no attention to the issue of value seems to have occurred in the field of aviation forecasting. Given the large financial impact that weather forecasts have on airline operations, particularly terminal aerodrome forecasts (TAFs), the topic would seem to deserve some recognition.

The results of this paper will show that considerable potential value is lost by the traditional method of providing weather information in TAFs in categorical form, that is, as a binary, yes–no product. The current rules do allow forecasters to use probabilities, for some elements like thunderstorms and fog. For example they can say ''PROB30'' for the occurrence, which means a 30% chance of occurrence. However, airlines are obliged by regulations to carry the full fuel requirement equivalent to a forecast of 100% confidence. So the use of the PROB30 is redundant, and the forecast is effec-

* Current affiliation: Weathernews Americas, Inc., Norman, Oklahoma.

Corresponding author address: Ross Keith, 14 Lawson St., Townsville 4812, Australia.
E-mail: Keiths14@bigpond.net.au

tively completely categorical. Forecasters know this, and also have some concept of the consequences of missed events. It will be shown that they adopt quite varying tactics, demonstrating a variety of attitudes to these consequences. Forecasters are also able to use temporal categorical variations to mean conditions when formulating TAFs. If the mean conditions are above the special lowest alternate minimum (SLAM), the term INTER (TEMPO) refers to periods of less than 30 (up to 60) min below the SLAM. It is shown below that use of these modifiers can be interpreted as crude probabilistic forecasting.

2. Experiment design

Forecasters at three Australian forecasting offices were asked to estimate their confidence, to the nearest 10%, that the weather at five different lead times will be below the SLAM for a particular aerodrome. The SLAM comprises values of cloud base and visibility, as well as weather type, and is the level used to determine fuel carriage for most passenger-carrying aircraft. The lead times are 1, 3, 6, 12, and 18 h. Forecasters formulated the probabilities at the same time they produced the routine issue of the TAF. Only routine issues of the TAF were tested. This was done so that the lead time–skill relationship was not skewed. Nonroutine amendments are usually issued to amend the TAF at short lead times, and so counting these would probably bias the skill in favor of the shorter lead times.

One purpose of the experiment was to demonstrate the use of the signal detection theory (SDT) model of forecasts for TAFs, and to ascertain whether the data can be fitted to the Gaussian model as used by Mason (1982). Another aim was to investigate differences in the forecast tactics between individual forecasters, and to demonstrate the effects of these differences on the financial outcome to airlines.

Other studies, most notably Mason (1982), have shown that probabilistic forecasts of elements like rain, storms, and temperature closely fit the SDT model. But in order to use Mason's cost–value model (I. M. Mason 2001, personal communication and described below), it was necessary to confirm that forecasters' probability estimates of poor weather at airports fit the Gaussian assumptions of the SDT model.

The trial data were provided by forecasters at the Victorian Regional Forecasting Centre (Vic RFC), the Sydney Airport Meteorological Unit (SAMU), and the Townsville Meteorological Office (TVL). All forecasters were volunteers. Data have been accumulated at Townsville since December 1999, at Vic RFC since March 2000, and at SAMU from about April 2000. Data have been analyzed up to September 2001.

3. Theory

a. Signal detection theory

The primary advantage of SDT is its ability to model the effect of forecasters' decision thresholds on accuracy. It provides a measure of a forecaster's acuity, or ability to discriminate between two signals, one just noise, and one noise plus signal. For a treatment of SDT as applicable to weather, the reader is referred to Mason (1982). Other interesting applications of SDT in meteorology are described in Harvey et al. (1992) and Levi (1985). Swets (1996) gives a full treatment of SDT in general, with most applications being in the medical and psychological fields.

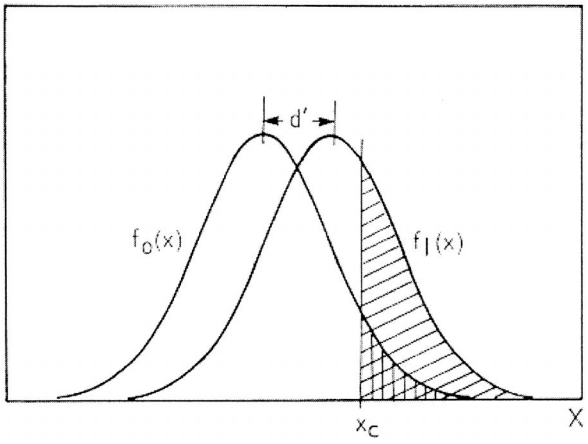


FIG. 1. Idealized probability distributions of the decision variable χ, $f_0(x)$ preceding nonoccurrence, and $f_1(x)$ preceding occurrence of the predictand. Here, χ_c represents the decision criterion. The area marked by vertical hatching indicates the probability of a false alarm and the area of diagonal hatching represents the probability of a hit. [From Mason (1982).]

The linchpin of SDT is the assumption that, prior to a decision, there are two overlapping probability distributions: the weight of evidence for the event occurring, and weight of evidence *against* the event occurring. This is illustrated in Fig. 1, with $f_1(x)$ representing the evidence for the event, and $f_0(x)$ the evidence against. The y axis is the weight of evidence, the x axis shows values of χ, the decision threshold. The area under the probability distribution for $f_1(x)$ to the right of χ_c, the critical decision threshold, is the probability of a hit (hit rate). Similarly, the area under the $f_0(x)$ curve to the right of χ_c is the probability of a false alarm (false alarm rate). If the forecaster becomes more conservative (i.e., his critical decision threshold χ_c decreases or moves to the left), both the hit rate and false alarm rate increase. Conversely, if the forecaster becomes more adventurous (i.e., less inclined to forecast the event), the false alarm rate decreases, but so too does the hit rate. Note that the separation of the means of the two distributions is denoted as d'. This parameter can be used as an index of skill as it defines the intrinsic ability to discriminate between the two distributions.

The form of the distributions in Fig. 1 is assumed to be Gaussian. Studies of human decision making under conditions that produce uncertainty have demonstrated Gaussian behavior. Mason (1982) showed this for weather forecasts. This behavior is demonstrated for the

data in this experiment below. The Gaussian assumption is also convenient in that the calculations fall out nicely.

The formal definition of hit rate, h, is Pr(Forecast = Yes | Event = Yes), that is, the probability that the event is forecast given that it happens. This is the same as the probability of detection (POD) as widely used in meteorology. Similarly false alarm rate, f, is Pr(Forecast = Yes | Event = No), that is, the probability that the event is forecast given that the event does not happen. Note that f is different from FAR, the false alarm ratio, a term encountered in meteorology and often confused with f. FAR in fact is Pr(Event = No | Forecast = Yes).

In the trial, forecasters were asked to indicate their confidence of below SLAM weather at each lead time to the nearest 10%, so that there are 11 different decision thresholds from 0% to 100%. The hit rate and false alarm rate can then be calculated for forecast probability greater than or equal to each of the 11 different decision thresholds. These are then plotted against one another to produce a ROC. By way of example, the ROCs produced in this experiment are shown later (Figs. 5–7).

The diagonal on the ROC is the line $h = f$. On the diagonal a forecaster has the same chance of a hit as a false alarm, and so the diagonal is defined as zero skill. A perfect forecast is indicated by a ROC from (0,0) to (0,1) to (1,1). So a reasonable skill measure is the area under the ROC curve. Mason (1982) calls this A_z once the Gaussian model is applied. Here, A_z has been shown to be independent of the climatological rate of occurrence (e.g., Mason 1989) and also, most importantly, *independent of the decision criterion* χ_c (e.g., Harvey et al. 1992). These are advantages of the SDT summary measure A_z. The dependence of some traditional skill scores on decision thresholds is described in Swets (1986, 1996), Mason (1989), and Harvey et al. (1992).

As mentioned in Harvey et al. (1992), A_z is a measure of potential forecast performance. As is demonstrated in their paper, and in this study, the economic value of the forecast depends on the decision criterion, χ_c, as well as on the accuracy. As a measure of skill, A_z qualifies as a *strictly proper* score. This means that it is not possible to optimize the score by hedging. Because conventional skill scores are, to varying degrees, dependent on decision threshold, a shrewd forecaster, after a little research and experimentation, could optimize his or her score by adopting a decision threshold that optimizes the particular score by which he or she is being assessed.

Returning to the Gaussian distribution of the signal detection model in Fig. 1, h and f can be expressed in terms of the location of the decision threshold, χ_c, on the x axis of the overlapping normal distributions. A more expansive treatment of this can be seen in McMillan and Creelman (1991), Swets (1996), or Mason (1982).

The likelihood ratio is defined as $\beta = f_1(x_c)/f_0(x_c)$. If the mean of f_0 is set to zero, the mean of f_1 becomes d', the separation of the means. If the variances of the two distributions are assumed to be equal, one can connect χ_c and d' as follows. From the formula of the normal distribution,

$$\beta = \exp[-0.5(\chi_c - d')^2]/\exp(-0.5\chi_c^2) \quad \text{and}$$

$$2\ln\beta = 2\chi_c d' - d'^2, \quad \text{so } \chi_c = \ln\beta/d' + d'/2.$$

From Mason (1982), the likelihood ratio β can be expressed through Bayes's theorem as a ratio of the odds of the probabilities of the event at $\chi = \chi_c$, p, to the odds of the climatological probability, p_c:

$$\beta = [p/(1 - p)]/[p_c/(1 - p_c)]$$

A link is thus available from forecast probability to hit rates and false alarm rates, using as variables p_c and d'. Again assuming equal variances of the two distributions, d' can readily be calculated from the data in the standard 2 × 2 contingency table of Table 1. Here, d' is just the difference between the normal deviates of h and f. As h is an estimate of Pr(Forecast = Yes | Event = Yes), $h = d/(d + b)$, and similarly $f = c/(c + a)$. The normal deviates of these can be calculated and subtracted to yield d'. The degree of validity of the assumption of equal variances in the context of this paper will be discussed below.

TABLE 1. A 2 × 2 contingency table of event forecasts and outcomes.

	Observed	
Forecast	No	Yes
No	a	b
Yes	c	d

b. Forecast value

In any forced choice, binary outcome (yes–no) forecast situation, the outcome can be summarized by the traditional 2 × 2 contingency, for example, Table 1. The outcomes, for the sake of intuitive understanding, are described as true positives (hits), true negatives (correct rejections), false negatives (misses), and false positives (false alarms).

Harvey et al. (1992) developed a relationship for expected value of a forecast, using the four conditional probabilities from the contingency table, and the value of each of the outcomes. They arrive at a relationship for the expected value (EV) of a forecast. Substituting terminology used in this paper for their terminology, the relationship is

$$\mathrm{EV} = hp_cV_{\mathrm{TP}} + (1 - p_c)fV_{\mathrm{FP}} + p_c(1 - h)V_{\mathrm{FN}} + (1 - p_c)(1 - f)V_{\mathrm{TN}}, \quad (1)$$

where p_c is the climatological rate of occurrence of the event, or Pr(Event = Yes), the Bayesian prior probability; h is hit rate; f is false alarm rate; and V_{TN} is the value of a true negative, V_{FN} the value of a false negative, V_{FP} the value of a false positive, and V_{TP} the value of a

true positive. Both V_{FN} and V_{FP} have negative values and are better called costs.

Equation (1) was derived independently by Mason (2001, personal communication). He uses the fact that perfect forecasts have $h = 1$ and $f = 0$, and so he simplified (1) to provide an expression for the expected cost of an imperfect forecast *with respect to a perfect forecast*:

$$\text{expected cost} = (1 - p_c)f(V_{TN} - V_{FP}) + p_c(1 - h)(V_{TP} - V_{FN}).$$

Mason defines $(V_{TN} - V_{FP})$ as the false alarm cost, the cost of incorrectly forecasting an event, and $(V_{TP} - V_{FN})$ as the miss cost, the cost of not forecasting an event. Making this substitution,

$$\text{expected cost} = (1 - p_c)f(\text{false alarm cost}) + p_c(1 - h)(\text{miss cost}). \quad (2)$$

Note that by definition, miss cost *does not* include any costs already incurred by a hit, or correct forecast of an event. In the context of a TAF, an aircraft may not be able to land regardless of whether the bad weather was or was not forecast. The miss cost is only that cost attributable to the event not being forecast over and above the cost accrued if it *had* been correctly forecast. False alarm cost is easier to ascribe, as there is no weather-related cost attributable to a correct rejection, and it is thus the cost of protective action. In the TAF situation, this would be the cost to carry extra fuel and perhaps lost payload.

Mason (2001, personal communication) has shown that if (2) is differentiated with respect to p, the forecast probability, and the result set equal to 0, the value of p(opt), the value of the optimal forecast probability that minimizes the expected cost, is defined thus:

$$p(\text{opt}) = \text{CR}/(1 + \text{CR}), \quad (3)$$

where CR is the cost ratio and equals the false alarm cost divided by the miss cost.

Mason's derivation of (3) goes as follows: differentiating (2) with respect to p and setting the result to zero gives

$$(1 - p_c) \cdot (df/dp)\text{CR} - p_c(dh/dp) = 0, \quad \text{so}$$

$$dh/df = \text{CR}(1 - p_c)/p_c.$$

It can been shown (e.g., Green and Swets 1974) that the slope of the ROC, dh/df, is the same as the likelihood ratio, β, at the corresponding value of χ_c. Substituting $[\beta/(p/(1 - p))]$ for $(1 - p_c)/p_c$ gives

$$p = p(\text{opt}) = \text{CR}/(1 + \text{CR}).$$

So, as one would logically expect, the optimum decision threshold is a function of the costs of the outcomes only. For a particular flight, the magnitude of the expected cost will be determined by the skill d' and p_c, but for a given d' and p_c the minimum value of expected cost will be at $p = p$(opt).

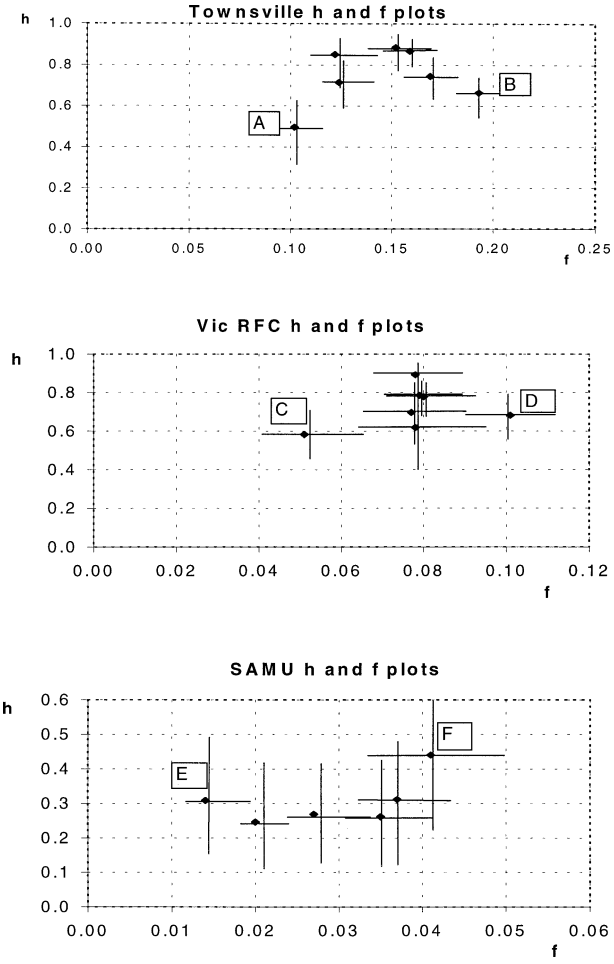


FIG. 2. Plots of hit rate vs false alarm rate for individual forecasters at Townsville, Vic RFC, and SAMU. Error bars are 95% confidence intervals.

4. Results and discussion

a. Differences between forecasters

Figure 2 shows plots of hit rate and false alarm rate for individual forecasters at Townsville, Vic RFC, and SAMU. These data are derived from the Bureau of Meteorology's automated TAF verification system and are based on 3–4 yr of verification data. The data are different from the experiment in this paper and were used in order to achieve better error statistics. They are shown merely to demonstrate the range of differences between individuals. The values of h and f shown are a composite of each whole hour of lead time up to 6 h; 95% confidence limits are included. The confidence intervals have been calculated from the formula quoted in Stephenson (2000).

While most forecasters aggregate around particular

(a)

TVL Forecaster A — Number of forecasts: 323

Distribution of INTER forecasts:

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	No. of times INTER forecast	No. of actual events
1 hr lead													
	0	6	11	11	0	0	3	0	0	0	0	31	11
3 hr lead													
	0	3	14	16	2	1	0	0	0	0	0	36	14
6 hr lead													
	0	4	17	18	2	1	0	0	0	0	0	42	11
12 hr lead													
	0	4	19	18	1	0	0	0	0	0	0	42	9
18 hr lead													
	0	4	14	14	1	1	0	0	0	0	0	34	9

Distribution of TEMPO forecasts

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	No. of times TEMPO forecast	No. of actual events
1 hr lead													
	0	0	2	3	0	1	2	0	4	1	0	13	11
3 hr lead													
	0	1	2	6	1	6	0	0	1	0	0	17	14
6 hr lead													
	0	2	5	7	2	5	0	0	0	0	0	21	11
12 hr lead													
	0	3	6	9	1	3	0	0	0	0	0	22	9
18 hr lead													
	0	3	3	8	0	3	0	0	0	0	0	17	9

(b)

TVL Forecaster B — Number of forecasts: 259

Distribution of INTER forecasts:

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	No. of times INTER forecast	No. of actual events
1 hr lead													
	0	1	1	4	3	3	0	0	0	0	0	12	5
3 hr lead													
	0	0	1	9	4	3	0	0	0	0	0	17	4
6 hr lead													
	0	0	4	13	2	1	0	0	0	0	0	20	8
12 hr lead													
	0	0	5	11	3	0	0	0	0	0	0	19	4
18 hr lead													
	0	0	6	9	1	0	0	0	0	0	0	16	5

Distribution of TEMPO forecasts

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	No. of times TEMPO forecast	No. of actual events
1 hr lead													
	0	0	1	1	2	2	5	2	0	0	0	13	5
3 hr lead													
	0	0	0	4	3	2	1	0	0	0	0	10	4
6 hr lead													
	0	0	1	2	6	0	1	0	0	0	0	10	8
12 hr lead													
	0	0	5	5	7	2	0	0	0	0	0	19	4
18 hr lead													
	0	0	5	6	5	2	0	0	0	0	0	18	5

FIG. 3. Number of forecasts of INTER and TEMPO, regardless of outcome, for three forecasters at Townsville.

regions in the graph, each plot shows forecasters with markedly different decision thresholds. In Fig. 2, forecasters A, C, and E show considerably less aversion to the risk of a miss than forecasters B, D, and F, respectively. For example, forecaster A has a smaller hit rate than B, and a much smaller false alarm rate. Therefore, A is operating at a higher decision threshold than B; that is, he or she is less cautious than B. The effect of the variation of the decision threshold on the cost outcome is described below.

In each set of data shown in Fig. 2, the difference in the decision threshold between the outliers is more marked than are differences in skill. This suggests that any difference in *value* may be caused more by varia-

(c)

TVL Forecaster C — Number of forecasts: 411

Distribution of INTER forecasts:

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	No. of times INTER forecast	No. of actual events
1 hr lead													
	0	1	4	9	4	0	0	0	1	1	0	20	17
3 hr lead													
	0	0	5	6	7	2	0	1	0	1	0	22	13
6 hr lead													
	1	1	7	12	4	1	1	0	1	0	0	28	18
12 hr lead													
	0	1	5	20	6	1	1	1	0	0	0	35	17
18 hr lead													
	0	0	7	14	5	3	3	0	0	0	0	32	12

Distribution of TEMPO forecasts

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	No. of times TEMPO forecast	No. of actual events
1 hr lead													
	0	0	3	0	3	5	0	2	5	5	4	27	13
3 hr lead													
	0	0	0	3	4	4	5	3	4	7	1	31	9
6 hr lead													
	0	0	2	6	0	9	8	6	4	0	0	35	12
12 hr lead													
	0	1	0	6	9	4	6	6	2	1	0	35	11
18 hr lead													
	0	1	0	12	9	7	2	2	3	0	0	36	8

FIG. 3. (*Continued*)

tions in decision threshold than by differences in skill, keeping in mind the usual shape of a line of equal skill on a ROC. The section above on forecast value shows how minimizing the cost of the uncertainty in the forecast depends on using an optimal decision threshold.

Climatological rates of below SLAM events at most Australian airports are typically around 0.02. As such, there is not adequate data from the experiment in this paper with which to carry out an analysis on individual forecasters. There are four entries daily for each TAF, and given that there can be up to eight or nine forecasters involved, there may be only 10 to 20 events for each forecaster for each lead time. To acquire enough data for analysis of individual forecasters would take several years. It is hoped to acquire enough of the probability forecasts over time at Townsville, where data acquisition continues, to enable ROC analysis on individuals. Nonetheless some conclusions can be made as to an individual's proclivity to forecast below SLAM weather more or less than others, simply by analyzing Pr(Forecast = Yes) for all confidence levels. This is irrespective of whether the event happened or not, and comparison assumes a reasonably constant frequency of occurrence among the group over the period. Only gross and obvious differences between individuals are discussed. This analysis is shown for three forecasters at Townsville in Fig. 3, and for two forecasters at Vic RFC in Fig. 4. There were not enough events forecast at SAMU to enable such an analysis.

At Townsville, ALT (alternate) is not forecast all that often, due to the climatology. Below minimum events are mostly precipitation induced and convective in nature so, unlike say fog, there are usually occasional breaks in the precipitation. So the analysis in Fig. 3 is confined to INTER and TEMPO forecasts. At Melbourne, however, most of the below SLAM weather is due to fog and low cloud, which tends to be persistent over a period of hours. Forecasts of TEMPO are not made very often because precipitation-induced events are usually brief. So the analysis in Fig. 4 is confined to INTER and ALT forecasts.

Data on the three individual forecasters from Townsville in Fig. 3 demonstrate considerable variation in their perceived confidence of below minimum conditions, and their inclination to forecast below minimum conditions. Each forecaster had issued well in excess of 200 forecasts. The total number of actual occurrences of below minimum conditions at all the target times is listed in the last column. Generally all three are less conservative at short lead times for both INTER and TEMPO forecasts; that is, the required confidence for the forecast of the event is higher at short lead times than longer lead times. Of interest is the large range of decision thresholds for forecaster C, especially for TEMPO forecasts, compared to the others. Both A and B rarely perceive a confidence greater than 50% at lead times greater than 1 h, whereas C for TEMPO forecasts exhibits a broad almost uniform spread of confidence across the whole range of probabilities. Forecaster C also tends to forecast TEMPO at a higher rate than the others, in line with his perceived confidence. By contrast, forecaster A is more inclined to forecast INTER conditions even though his perceived confidence is much the same, or even a little less, than the others.

The clustering of the decision thresholds around certain percentages for forecasting INTER and TEMPO can

(a)

VRO Forecaster A — Number of forecasts: 247

Distribution of INTER forecasts:

1 hr lead	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	No. of times INTER forecast	No. of actual events
	0	0	2	3	6	0	0	0	0	0	0	11	5
3 hr lead													
	0	0	3	5	2	1	0	0	0	0	0	11	4
6 hr lead													
	0	0	4	6	2	0	0	0	0	0	0	12	9
12 hr lead													
	0	1	3	8	5	1	0	0	0	0	0	18	11
18 hr lead													
	0	1	3	9	6	1	0	0	0	0	0	20	9

Distribution of Alternate forecasts

1 hr lead	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	No. of times INTER forecast	No. of actual events
	0	0	0	1	0	0	1	2	1	3	2	10	5
3 hr lead													
	0	0	0	1	0	0	1	3	3	2	2	12	5
6 hr lead													
	0	0	0	1	0	0	1	2	4	3	2	13	9
12 hr lead													
	0	0	1	2	0	1	1	3	4	1	1	14	11
18 hr lead													
	0	0	0	2	0	0	1	3	3	2	0	11	9

VRO Forecaster B — Number of Forecasts: 267

(b)

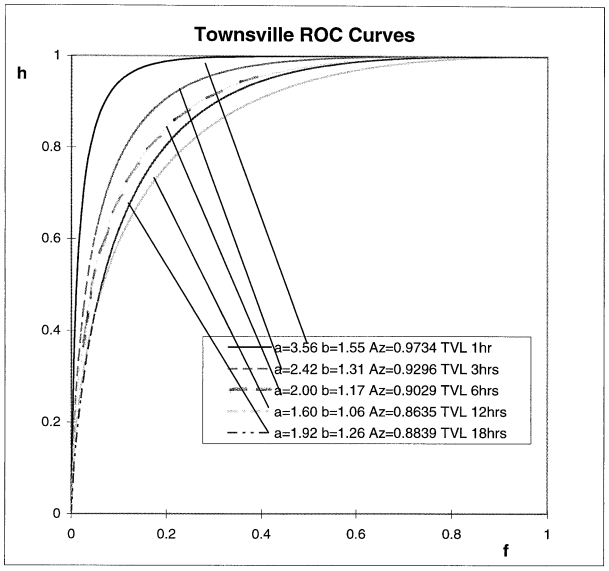
Distribution of INTER forecasts:

1 hr lead	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	No. of times INTER forecast	No. of actual events
	0	3	2	3	2	0	0	0	0	0	0	10	4
3 hr lead													
	0	3	4	3	3	1	0	0	0	0	0	14	8
6 hr lead													
	0	4	4	5	3	1	0	0	0	0	0	17	7
12 hr lead													
	0	3	2	4	3	0	0	0	0	0	0	12	9
18 hr lead													
	0	6	4	3	1	0	0	0	0	0	0	14	8

Distribution of Alternate forecasts

1 hr lead	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	No. of times ALT forecast	No. of actual events
	0	0	0	1	2	3	1	0	2	1	0	10	4
3 hr lead													
	0	0	0	1	1	3	1	2	1	0	0	9	8
6 hr lead													
	0	0	1	1	2	6	1	1	0	0	0	12	7
12 hr lead													
	0	1	1	2	3	1	2	2	0	0	0	12	9
18 hr lead													
	0	0	0	1	1	5	1	0	0	0	0	8	8

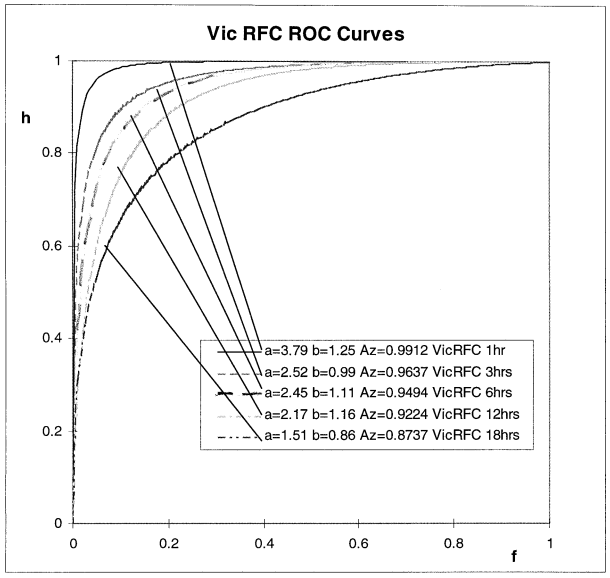
FIG. 4. Number of forecasts of INTER and ALT, regardless of outcome, for two forecasters at Vic RFC.



95% confidence limits for A_z

1 hour lead: $A_z = 0.9734$ (0.9613, 0.9821)
3 hour lead: $A_z = 0.9296$ (0.9049, 0.9490)
6 hour lead: $A_z = 0.9029$ (0.8682, 0.9303)
12 hour lead: $A_z = 0.8635$ (0.8188, 0.9000)
18 hour lead: $A_z = 0.8839$ (0.8424, 0.9170)

FIG. 5. ROCs for Townsville, with 95% confidence limits for A_z. Here, a is the y intercept of the plot of normal deviates of h and f on binormal axes, and b is the slope of a best fit straight line of the plot of the normal deviates.



95% confidence limits for A_z

1 hour lead: $A_z = 0.9912$ (0.9809, 0.9963)
3 hour lead: $A_z = 0.9637$ (0.9321, 0.9821)
6 hour lead: $A_z = 0.9494$ (0.9176, 0.9706)
12 hour lead: $A_z = 0.9224$ (0.8849, 0.9498)
18 hour lead: $A_z = 0.8551$ (0.7924, 0.9036)

FIG. 6. As in Fig. 5 but for ROCs for Vic RFC.

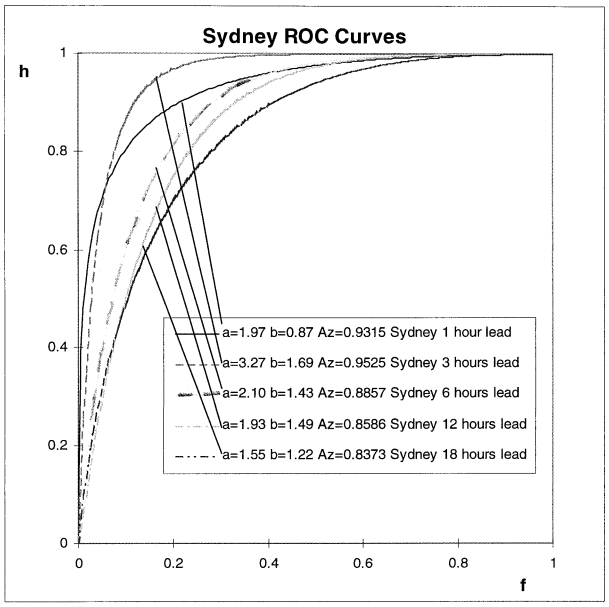
also be interpreted as the forecasters using these temporal variations as de facto probabilities. At Townsville, INTER forecasts cluster around 20%–30%, and TEMPO forecasts around 30%–40%. As forecasters are required to express the forecast in a categorical manner, above or below the airport minimum, they can express their confidence with INTER or TEMPO. For example, if forecaster C thinks there is a 30% chance of below minimum conditions 6 h ahead, he will generally forecast INTER conditions, but if he thinks there is 60% chance, he will generally forecast TEMPO. Discussions with Townsville forecasters generally revealed a strong tendency to forecast TEMPO if they thought thunderstorms or heavy showers would be frequent and widespread, and INTER if the thunderstorms would be isolated. Logically, if a forecaster thinks the poor weather could last up to 60 min, and so forecasts TEMPO, he or she believes there is more chance of below minimum conditions at a particular time than if he or she thinks the poor weather will last only up to 30 min; and forecasts INTER. So the temporal variations of INTER and TEMPO can also be interpreted as crude probability forecasts.

Referring now to Fig. 4 and the analysis for two forecasters at Vic RFC, there is a significant difference in the percentage confidence levels for forecasts of ALT between the two. One would expect the distribution for ALT forecasts to be bimodal. A forecast of a probability of 30% or 40% of fog or thunderstorms (phenomena that can preclude landing) is considered by the safety regulators as an alternate forecast. Forecaster A exhibits this bimodality, whereas the distribution for forecaster B is clustered about 50%. Forecaster B shows generally much higher probabilities for ALT forecasts. For both forecasters, the higher probabilities for ALT forecasts compared to INTER forecasts are very apparent.

The above discussion of the differences between forecasters is not meant to imply any specific effect on the economics of airline operations. The only purpose of this analysis is to demonstrate that there are obvious differences in the approach taken by different forecasters.

b. ROCs

Figures 5–7 show ROCs for Townsville, Vic RFC, and SAMU, respectively. These are maximum likelihood best fit curves based on the Gaussian model. They show A_z values generally decreasing with lead time. Exceptions are two small and insignificant reversals, one at Townsville between 12- and 18-h lead time, and the other between 1 and 3 h at SAMU. Both reversals are within the 95% confidence limits. The ROCs for SAMU are based on a number of events insufficient to make any firm conclusions about variation of skill with lead time. Data acquisition is continuing at SAMU. The 95% confidence limits for SAMU are very broad, and these ROCs are only presented for the sake of completeness.



95% confidence limits for A_z

1 hour lead: $A_z = 0.9315$ (0.7744, 0.9868)
3 hour lead: $A_z = 0.9525$ (0.8884, 0.9831)
6 hour lead: $A_z = 0.8857$ (0.7488, 0.9588)
12 hour lead: $A_z = 0.8587$ (0.7217, 0.9407)
18 hour lead: $A_z = 0.8373$ (0.6459, 0.9444)

FIG. 7. As in Fig. 5 but for ROCs for SAMU.

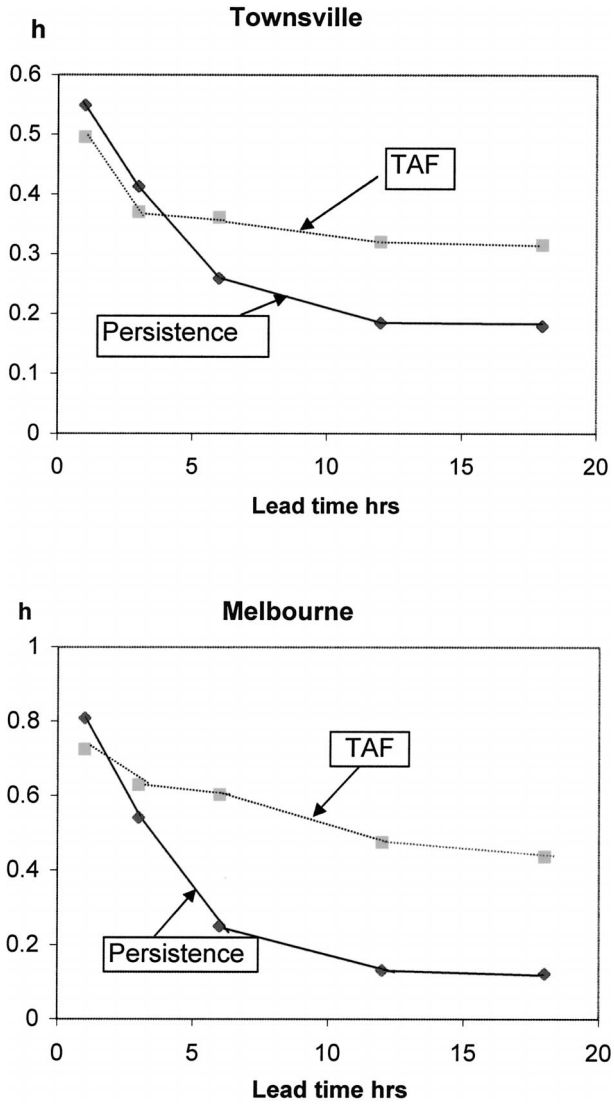


FIG. 8. Lead time vs hit rate for Townsville and Melbourne.

For Townsville and Vic RFC, however, it can be stated with reasonable confidence that skill decreases with lead time.

Figure 8 shows the relationship between hit rate and lead time for Townsville and Vic RFC (Melbourne) for both the forecasts and persistence. The hit rate for the forecast is calculated from the best fit ROC at the same false alarm rate calculated for persistence. The (h, f) pairs for persistence plot toward the bottom left of the ROCs, where the hit rate varies strongly with small changes in false alarm rate. Fixing the value of f thus enables a sensitive comparison of h. Persistence in this context means that if the initial conditions at the airport are below (above) the alternate minimum, the persistence forecast for all lead times is for below (above) the alternate minimum.

The persistence ratings on the ROCs for the Townsville and Melbourne data are interesting when compared to the findings by Harvey et al. (1992). Using the same method, they found that forecasts of convection at Stapleton International Airport, Denver, Colorado, with lead times of 1 h and less failed to match persistence. The data in that study and the current experiment are quite similar. Both involve short-term forecasts at airports and both have operational significance to aircraft operations. The results of the two experiments suggest that forecasts at lead times of less than about 3 h cannot beat persistence. The performance at a lead time of 3 h at Townsville fails to match persistence. Forecast skill in the Tropics at all timescales is known to be inferior to that in higher latitudes. This is of course due to the sporadic nature of events. At Vic RFC the performance at the 3-h lead time is a little better than persistence.

Figures 9–11 are plots of the normal deviates of hit rate and false alarm rate, for each lead time, for Townsville, Vic RFC, and SAMU. Note that there is no plot for SAMU for 12- and 18-h lead times. Due to the rare nature of the events at Sydney, there were very few high confidence estimates at the longer lead times. Each point on the graphs is the value of the normal deviates of an (h, f) pair at a different decision threshold. The degree of linearity of these plots is a measure of the validity of the assumption of normality of the original distributions of hit rate and false alarm rate. Further, it can be shown (e.g., Green and Swets 1974) that the slope of the plot of normal deviates of h and f is equal to the

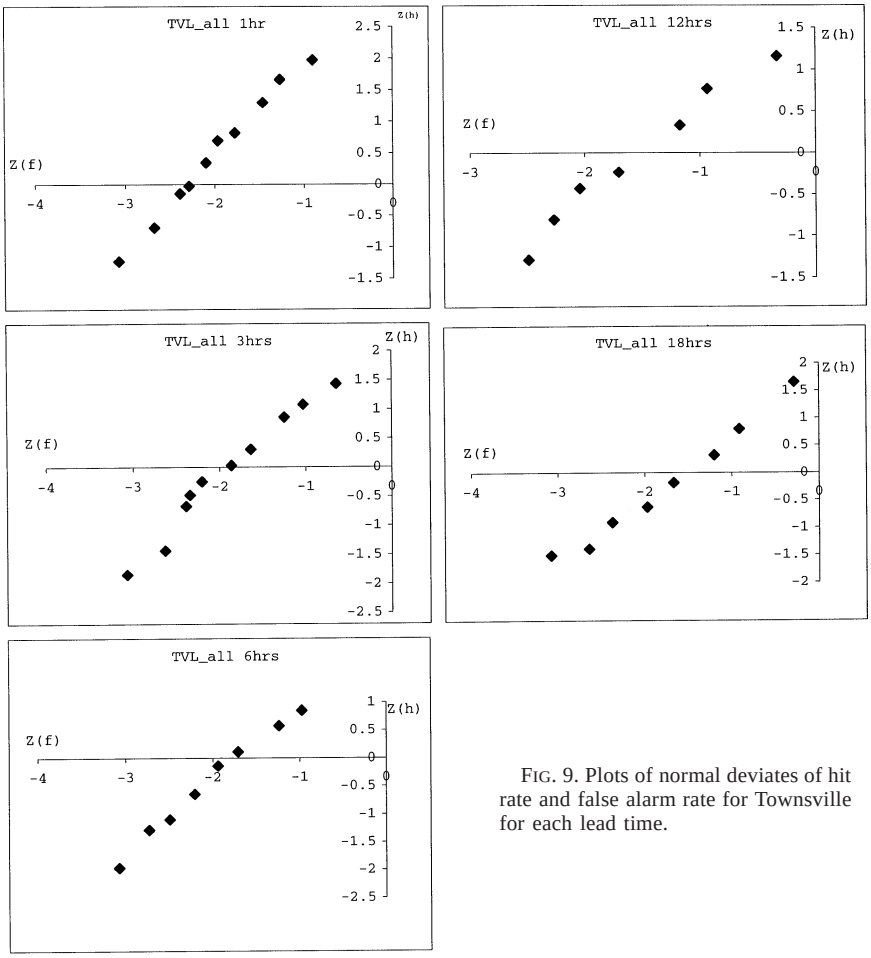


FIG. 9. Plots of normal deviates of hit rate and false alarm rate for Townsville for each lead time.

ratio of the standard deviations of the two distributions for and against the event. The proximity of this slope to unity gives a measure of the appropriateness of using d' as a measure of accuracy. Values of the slope are given as b in the legend of Figs. 6–8. As can be seen, the slope is between 0.99 and 1.31 for Townsville and Vic RFC for lead times of 3, 6, and 12 h. These lead times cover the period that is generally the most important for aviation operations. Use of d' is computationally easier and more pleasing than using A_z, which is a more robust measure of accuracy. It is not intended that the cost results derived using d' be used as an exact figure, but more as a neat illustration of the relative benefit of using probabilistic TAFs.

c. Reliability diagrams

Reliability diagrams for Townsville, Vic RFC, and SAMU are shown at Figs. 12–14, respectively. The reliability diagrams for Townsville and Vic RFC show clear overforecasting. The pattern of the overforecasting is similar to that observed by Murphy and Daan (1984). The poor reliability at SAMU is of concern. The rare nature of the events is undoubtedly a factor at Sydney. It is difficult to give high confidence to a rare event, especially at the longer lead times. The high impact of missed events at Sydney is probably also a factor.

It is tempting to use the knowledge of the past bias of individuals to recalibrate future forecasts toward greater reliability. Harvey et al. (1992), in a study of very short lead time ($<=$60 min) forecasts of storms, split the data into high and low activity days and showed the effect of stress on A_z and χ_c. The split was a simple median division. They found that on high activity days, the decision threshold was more cautious, that is, a (lower) χ_c. The study suggests that the forecasters' χ_c is not constant, and will vary depending on what Harvey et al. call stress. The factor causing the variation of χ_c to a more cautious value is possibly the forecasters' increased perception of a risk of adverse consequences when, in this case, the forecasters have a higher expectation of storms impacting on aircraft operations. Large amounts of data on individuals would be required to determine how their decision threshold varies with the weather situation, time, mood, etc. Another impediment to recalibration of probabilities produced by forecasters is the fact that, with TAFs, the events are usually infrequent. It would take several years to capture enough

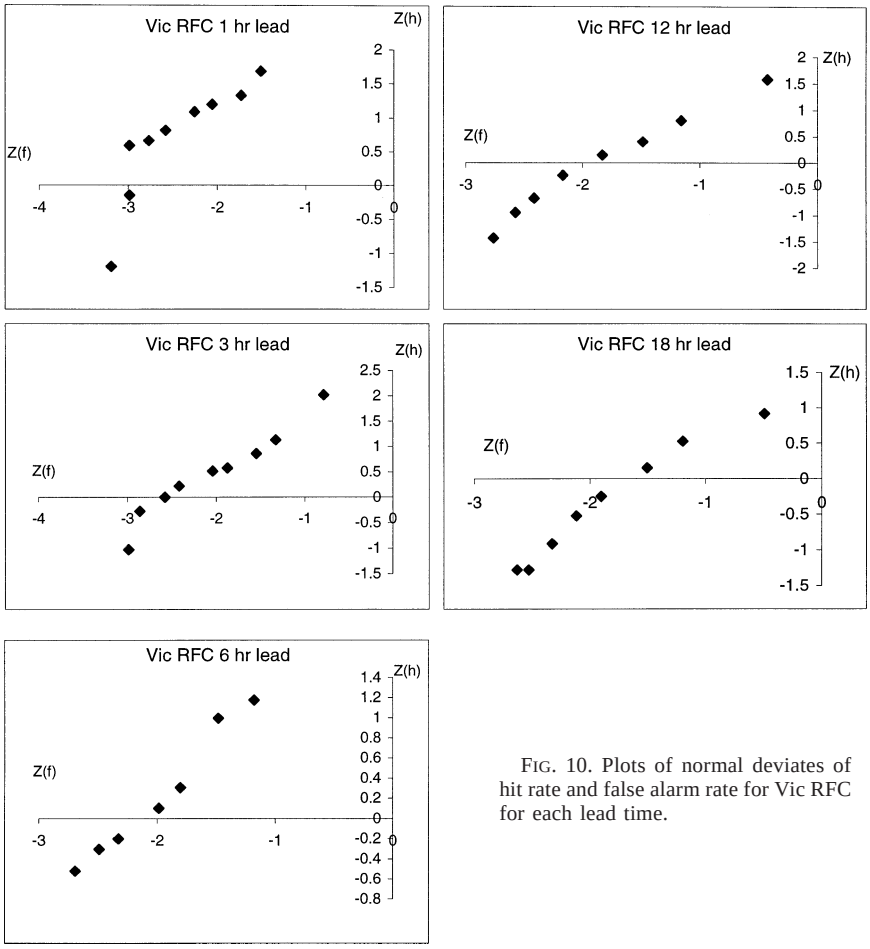


FIG. 10. Plots of normal deviates of hit rate and false alarm rate for Vic RFC for each lead time.

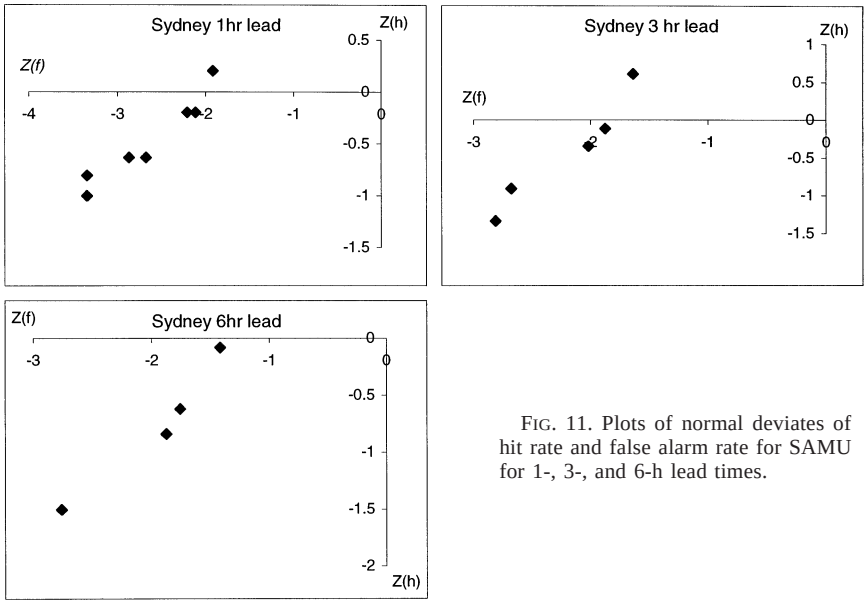


FIG. 11. Plots of normal deviates of hit rate and false alarm rate for SAMU for 1-, 3-, and 6-h lead times.

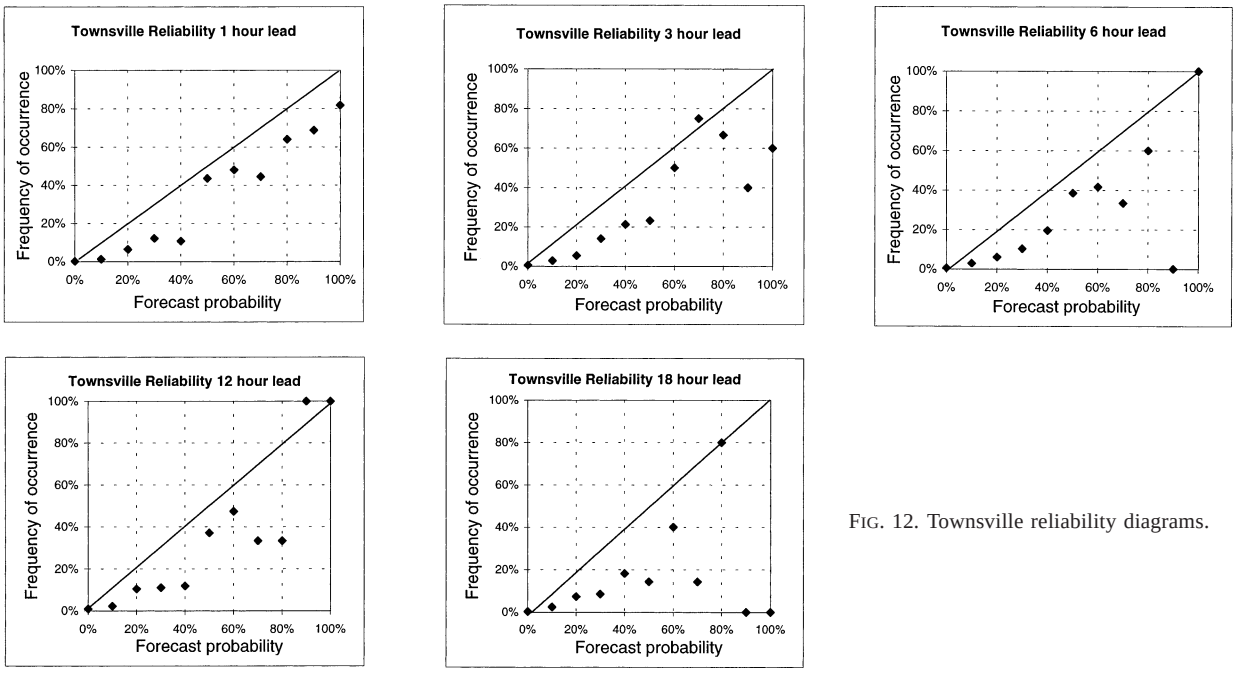


FIG. 12. Townsville reliability diagrams.

data to be able to recalibrate individual forecasters. By that time the forecasters would probably have transferred or retired. Murphy and Daan (1984) showed that, even with minimal feedback, forecasters can learn to improve their own reliability. Recalibration of automated forecasts produced by statistical or numerical methods would be less of a problem.

5. Cost analysis

Costs were provided by Qantas Airways for a flight from Singapore to Melbourne. The actual figures are commercially confidential, but the methodology for calculating the false alarm cost and miss cost is described using symbols for the dollar amounts.

When the flight plan is done prior to the flight leaving Singapore, the Melbourne TAF forecasts either alternate (ALT) conditions, INTER or TEMPO deteriorations in the weather, or good conditions above the alternate minima. For the sake of the current exercise, INTER and TEMPO forecasts are ignored. INTER forecasts (30 min of holding fuel) are absorbed into the company's reserve, and at Melbourne TEMPO is forecast only rarely. If the Melbourne TAF forecasts below alternate conditions, the aircraft takes on sufficient fuel to make an approach into Melbourne, and then to fly back to Adelaide if it cannot land at Melbourne. Adelaide is near the track about 300 mi NW of Melbourne. In this case the pilot will usually make an approach into Melbourne, knowing he or she has enough fuel to abort and fly back to Adelaide. If the Melbourne TAF forecasts conditions above the alternate minima, the pilot usually flies on to Melbourne and lands. However if the weather at Melbourne is below the alternate minima when the flight reaches Tailem Bend (TBD), a point on track abeam Adelaide, the pilot will divert to Adelaide.

The costs required in order to calculate the false alarm cost and miss cost are shown in Table 2. The false alarm cost is readily available from this information, being the ''cost to carry'' the extra diversion fuel of $\$C$. Calculation of the miss cost is more complex. Remember that the miss cost is that cost caused by a diversion *over and above that cost accrued if the below minimum weather was correctly forecast,* that is, above the cost of a hit. The cost of a hit depends on whether the pilot can actually land at Melbourne when the weather is below the alternate minimum. If the pilot can land safely at Melbourne when the forecast is a hit, then there is no extra cost over the false alarm cost of $\$C$. However if he cannot land, then the aircraft must fly to Adelaide, and return to Melbourne when able to land there. The cost of the diversion is $\$(C + H + E + F)$ minus $\$I$ for fuel that is not used. Then the question arises: what proportion of the time that the weather is below the alternate minimum can the pilot still land?

Every airport has a level below the alternate minimum to which a pilot may descend, at which time he or she aborts if visual reference of the airstrip has not been attained. This level is called the instrument landing system (ILS) minimum. Thirty years of synoptic observations were analyzed for the Melbourne airport. It was found that of the number of occasions that the alternate minimum was breached, 28% of these also breached the ILS minimum, thus precluding landing. So the calculation of hit cost proceeds thus:

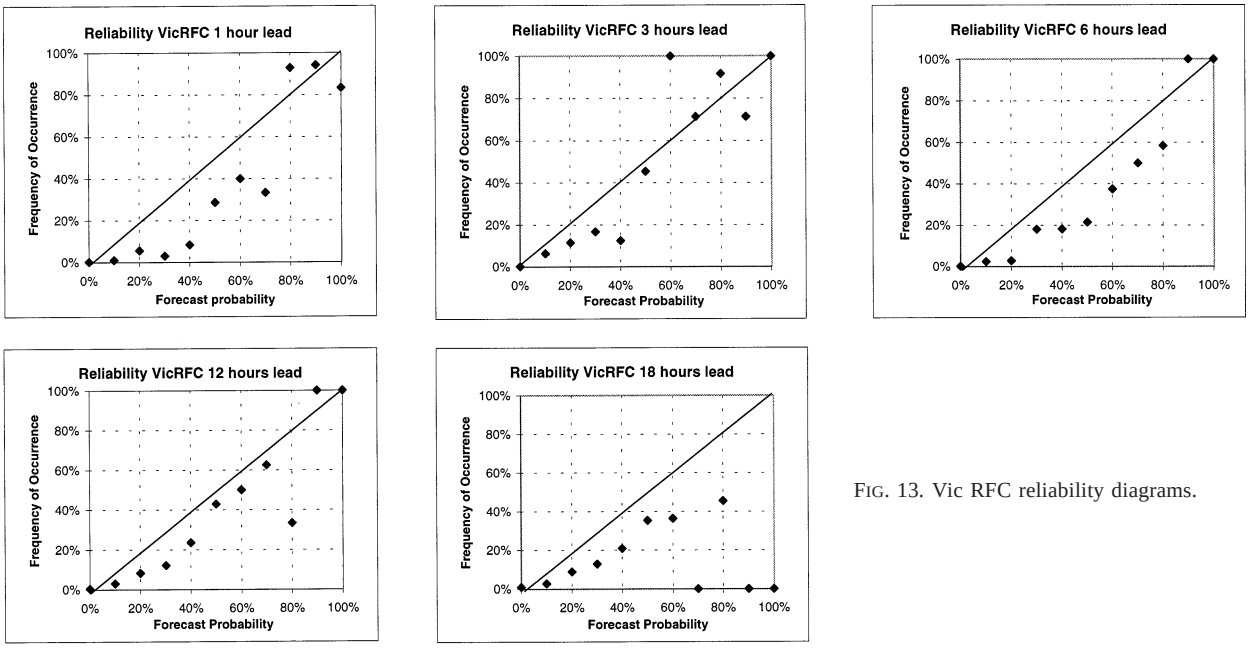


FIG. 13. Vic RFC reliability diagrams.

average hit cost

$$= (0.72 \times \$C) + [0.28 \times \$(C + H + E + F - I)] \quad \text{and}$$

the average miss cost

$$= \$(C + H + E + F - I) - \text{average hit cost.}$$

Inputting actual dollar values,

$$\text{CR} = \text{false alarm cost/miss cost} = 0.132, \quad \text{and}$$

$$p(\text{opt}) = \text{CR}/(1 + \text{CR}) = 0.117.$$

From the Bureau of Meteorology's TAF verification, a typical value of the (h, f) pair is (0.75, 0.12) for Vic RFC forecasters at a lead time of 6 h. Using the assumption of equal variances, $p_c = 0.02$ and $d' = 2.1$ (typical skill measured at Vic RFC in the experiment at

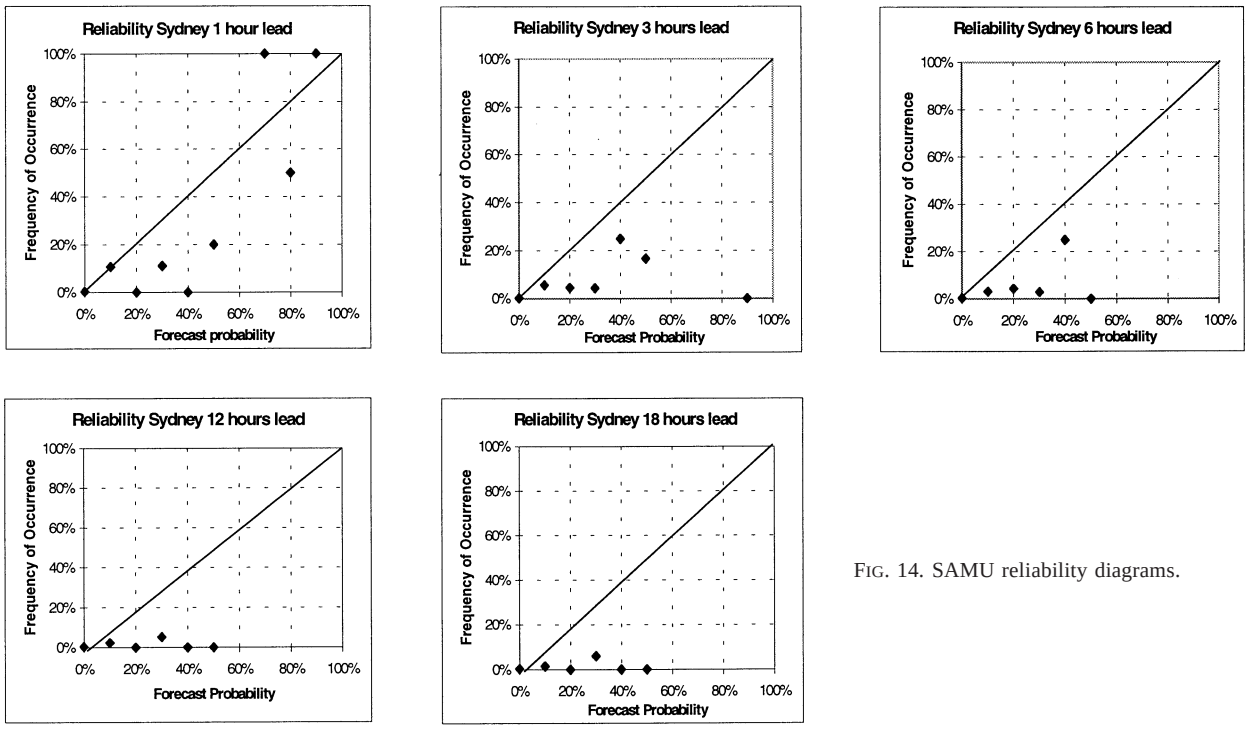


FIG. 14. SAMU reliability diagrams.

TABLE 2. Relevant costs for flights from Singapore to Melbourne.

Cost to carry diversion fuel	$\$C$
Cost to divert from Melbourne to Adelaide	$\$E$
Landing and handling fees at Adelaide	$\$F$
Cost of fuel flying Adelaide to Melbourne	$\$H$
Cost of unused fuel TBD into Melbourne	$\$I$

6- and 12-h lead time), the closest value of the forecast probability that produces (h, f) values nearest (0.75, 0.12) is about 0.02. So one can say that the forecasters at Vic RFC are operating with an effective average probability of about 0.02 as their decision threshold for this forecast.

Refer back now to (2) and use the false alarm cost and miss cost calculated for this flight. Figure 15 is a plot of cost versus decision threshold for this flight. At a decision threshold probability of 0.02, the cost of the uncertainty in this forecasts is \$231. If the forecast was reliably made at the optimum decision threshold, 0.117, the cost would have been \$128. So a perfectly reliable forecast of the probability of below alternate minimum conditions would, in the long run and at the same skill, save about \$103 per flight for the Singapore to Melbourne route. This is about 45% of the total cost of the uncertainty of the forecast.

Consider now the reliability diagrams for Vic RFC in Fig. 13. Assume a forecaster at Vic RFC forecast at a decision threshold of 0.117 for the Singapore to Melbourne flight. From the reliability diagrams, he or she would expect roughly 0.07 as the frequency of occurrence in the long run. This can then be considered as the effective decision threshold. A decision threshold of 0.07 leads to a cost of \$135. So for this flight, even the moderately reliable forecasts as currently produced would provide most of the savings (41%) gained by perfectly reliable TAFs (45%).

Figure 16 shows, for the same flight, how the modeled cost of the forecast varies with d', the index of skill. The two graphs are for the optimum decision probability of 0.117, and for the estimated actual decision threshold 0.02. The large difference in cost outcome for the two different decision probabilities is obvious, especially at low skill. Note that if the decision threshold is optimal, a decrease in skill does not matter all that much in economic terms. Therefore one could suggest that using near-optimal decision thresholds is more important than increasing skill, especially for low skill forecasts. The optimal decision threshold is solely a function of the operating costs of a particular flight and, thus, varies between flights. So reliable estimation of the probability of occurrence of the event must be applied to each flight in order to optimize savings, and the cost parameters for each flight used to determine whether extra fuel is required.

Note that the method of calculating the miss cost uses a quite crude climatology. The factor of 0.28, being the ratio of the time conditions are below the ILS minimum

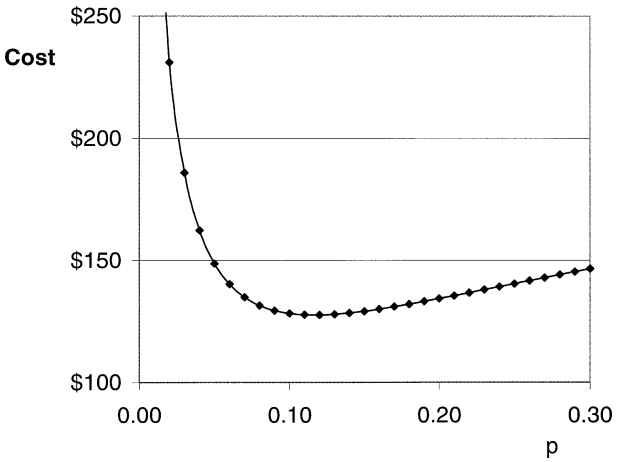


FIG. 15. Cost vs decision threshold for a Singapore to Melbourne flight. FAC = \$1390, MC = \$10 535, $d' = 2.1$, and $p_c = 0.02$.

to that below the alternate minimum, is for all times of the day and months. A superior value for the miss cost, and thus CR and p(opt), could be obtained if there were adequate data to determine a matrix of these factors for different times and months. This would undoubtedly increase the potential for savings by using probabilities. Furthermore, if the conditional probability of weather below the ILS minimum *given* that the weather is below the alternate minimum could be forecast with a skill better than climatology, a more accurate p(opt) value could, on average, be calculated on a case-by-case basis.

Figures 17 and 18 are the same two graphs for a shorter flight of duration about 2 h, between Brisbane and Townsville. The false alarm cost is \$200, the miss cost \$4800, $p_c = 0.01$, and the measured d' for Townsville forecasters was 2.0 at 3-h lead time. The measured average decision threshold for the group was 0.015, and taking into account the reliability diagram, the ''effective'' threshold would be about 0.01, even more con-

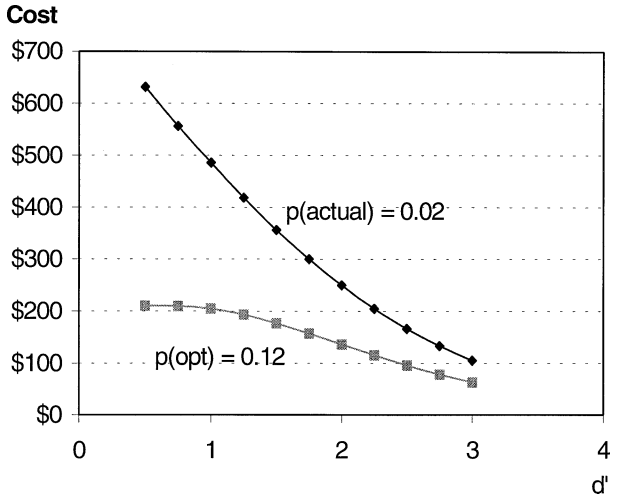


FIG. 16. Cost vs skill for a Singapore to Melbourne flight.

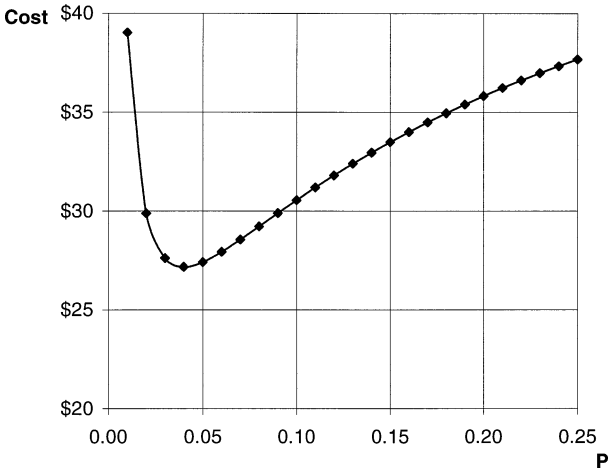


FIG. 17. Cost vs decision threshold for a Brisbane to Townsville flight. FAC = \$200, MC = \$4800, $d' = 2.0$, and $p_c = 0.01$.

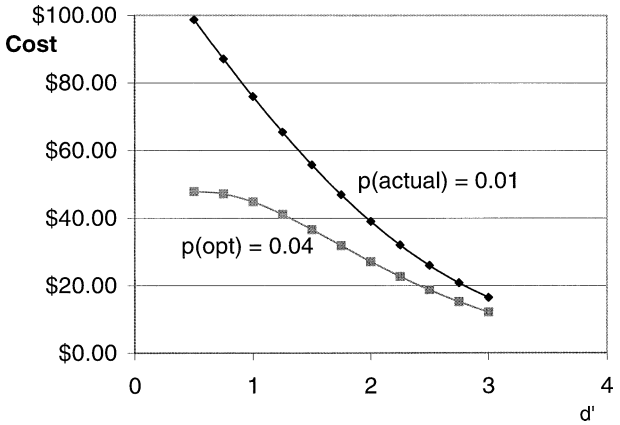


FIG. 18. Cost vs skill for a Brisbane to Townsville flight.

servative than that for Melbourne forecasters. At this threshold, the cost of the errors in this forecast is \$38, and at the optimal decision threshold of 0.04 the cost is \$27. So, in relative terms, the savings accrued by the use of probabilities is less for the shorter flight than for the longer, international flight. This is undoubtedly because of the relatively high false alarm cost, that is, the cost of carrying fuel over a long distance unnecessarily. Miss costs involve significant components that are fixed and do not depend on route length. Another interesting comparison between the two flights is the effect of a more adventurous (higher) decision threshold. The cost rises much more quickly from the minimum for the shorter flight as the decision threshold rises than for the long flight. In Fig. 17, if the decision threshold rises above 0.25, the cost savings accrued by using probabilities are lost. So the ''sweet spot'' for savings using probabilities is less than for the long flight.

6. Forecast system design

Meteorologists are judged on the outcome of forecasts, so it is reasonable to assume that they want to achieve the highest possible accuracy. On the other hand the desired outcome for commercial users of the forecast is to use the forecast to optimize the outcome of their particular enterprise.

Brown and Murphy (1987) studied the particular case of fire weather forecasts of credible interval temperature ranges, relative humidity, and wind speed. They found bias in the forecast of fire weather elements *toward* values that produced more caution, that is higher potential fire rate of spread. At higher wind speeds and lower relative humidities, forecasters are extremely cautious with their probability estimates. To quote from the abstract of that paper: ''The forecasts also exhibit modest but consistent biases which suggest that the forecasters are influenced by the impacts of the relevant events on fire behavior. These results underscore the need for probabilistic fire-weather forecasts.''

On the other hand, Winkler and Murphy (1979) and Murphy et al. (1989) looked at the reliability of credible interval temperature forecasts. Murphy and Winkler (1992), Hamill and Wilks (1995), and Murphy and Winkler (1976) studied probability of precipitation and temperature interval probability forecasts. A common element in all these studies was that there were no apparent safety or economic consequences to the estimates of probability. In all these studies the estimates were quite reliable; that is, the observed frequency of events closely resembled the forecast probability.

So it seems that forecasters can generally formulate reliable estimates of probability of occurrence *when the outcomes do not have known consequences that affect the decision.* On the other hand, if the outcomes have a safety connotation, or if a particular outcome can impact adversely on the forecaster, considerable overforecasting can occur. It is interesting that forecasters rarely, in fact never to the author's knowledge, have to write a report on a forecast error due to a false alarm on a TAF. However a missed event often triggers a ''please explain.''

Much has been written in the meteorological literature about probabilistic forecasting, in particular about its inherent value. Murphy (1977) discussed the relative value of climatological, categorical, probabilistic, and perfect forecasts. Murphy expressed forecast value as a function of the cost–loss ratio. He demonstrated that the value of the probabilistic forecasts was the same as the categorical forecasts when the cutoff probability used to define the categorical forecasts was about equal to the cost–loss ratio. At other values of the cost–loss ratio, the value of the probabilistic forecasts was greater. Mason (2001, personal communication) pointed out that the ratio CR/(1 + CR) is equivalent to the traditional cost–loss ratio. Returning to the earlier nomenclature for forecast value, $V_{TN} = 0$, $V_{FN} = L$, $V_{FP} = C$, and $V_{TP} = C$, where L is the loss due to an event and C is the cost of protection. By definition CR = $(0 - C)/(C -$

L), and the optimal decision threshold, $p(\text{opt}) = \text{CR}/(1 + \text{CR}) = C/L$, the cost–loss ratio. So the decision threshold that maximizes economic value is seen to be the same in the signal detection environment as that determined originally by authors in earlier papers.

Over the years there have been some attempts to cope with uncertainty in the TAF. Forecasters can specify in a TAF the probability of occurrence of fog and thunderstorms, these being elements that will probably preclude landing. The problem with this methodology is that the probability is ignored operationally, and *legally* airlines are bound to carry the extra fuel irrespective of the probability estimate. Several years ago, in order to correct what was perceived as excessive fuel usage in the United States and Europe where there are many suitable alternates, the International Civil Aviation Organization changed the rules and forecasters were instructed to mention only probabilities of occurrence of 30% and above on TAFs. In other words, the system was changed in a way that attempted to raise forecasters' decision threshold. In Australia, for example, many cautious forecasters in tropical areas were often forecasting 10% chance of thunderstorms, which was often a realistic estimate given the isolated and sporadic nature of these thunderstorms. When the instructions were changed, most forecasters just changed 10% to 30% (personal communication and experience!), *because the perceived consequences had not changed.*

As this change naturally tended to cause more misses, Qantas and the Bureau of Meteorology initiated the Code Grey system. This forecast is issued for major Australian airports during the afternoon, valid overnight and the next morning. Its purpose is to alert airlines to even the slightest chance of fog, even if this is only 1%–2%. As the miss cost for Quantas' international flights is high, they then carry fuel over and above that required by the categorical TAF. So Qantas realizes the value in the knowledge of a small probability of a significant event. If forecasters really operated on a decision threshold of over 0.30, the results would obviously be highly suboptimal. But this is what ''the system'' expects of them, and so they are caught in something of a dilemna.

It can readily be seen that the current TAF system is the result of a legalistic evolution to a system of categorical forecasting. The system takes no account of the probabilistic nature of forecasters' thought processes, in particular the variable decision threshold that is driven by consequences and individual forecaster's attitudes to those consequences.

7. Conclusions

Considerable economic benefit through lower fuel usage is potentially available to airlines if TAFs were expressed as estimated probabilities of breaching the alternate minimum. This would enable airlines to unlock the value in forecasters' ability to provide reasonably reliable estimates of the probability of occurrence of these events. The amount of benefit would depend on three main factors:

1) The ability of airlines to accurately specify the miss cost and false alarm cost for *every* flight. The decision process for the two flights given as examples is a straightforward and simple example. Many flights have multiple possible alternate destinations and PNR decision points. An example of the possible complications is a flight from Tokyo to Perth, Australia, operated by Qantas with a Boeing 767 (B767). This flight is near the limit of endurance for a B767. If the Perth TAF has an alternate requirement *and* the flight is carrying a full payload, the pilot must land at Darwin to take on extra fuel. So the false alarm cost increases dramatically due to payload considerations. The effect is to raise $p(\text{opt})$ from 0.02 to around 0.20, an order of magnitude greater. While many flights are simple, single-point decisions like the two examples, others are more complex, and considerable effort would be required by airlines to specify the full suite of costs.
2) The willingness of regulators to allow airlines to plan fuel requirements from a probabilistic forecast. For this to happen, it would be necessary to prove that probabilistic TAFs provide a commercial benefit to airlines and do not impact negatively on safety. To this end, a trial is commencing, in conjunction with Qantas and American Airlines. Costs will be developed for several flights from each airline. About an hour before departure, forecasters will estimate the probability of weather being below the alternate minimum at arrival time. A comparison will be made of the economic outcomes of this method of planning fuel and the traditional categorical TAF.
3) The ability of forecasters to achieve reliability with their probability estimates. The study in Murphy and Daan (1984) shows that forecasters can learn and that their probability estimates can become more reliable. The results of their experiment suggest that even better changes in reliability could be obtained with more frequent feedback to forecasters.

Acknowledgments. The author wishes to acknowledge the considerable assistance provided by Ian Mason, who pioneered the application of signal detection techniques to forecast verification. His extension of this work to the accuracy–value relationship will be the subject of a paper in the *Australian Meteorological Magazine* in the near future. Helpful critiques from Dr. Greg Connor (BoM Townsville office), Dr. Roger Atkinson (BoM Tindal office), and Dr. Warwick Grace (BoM Adelaide office) assisted with the content and presentation. Ted Williams of the aviation program office in the Bureau's Head Office provided data from the TAF verification scheme and also reviewed an early draft. Numerous meteorologists in the Bureau's Melbourne, Sydney, Towns-

ville, Brisbane, Adelaide, and Darwin offices input forecast data over one–two years, and their persistence as volunteers is greatly appreciated. Associate Professor Danny Coomans and Professor Mal Heron from James Cook University assisted at various times with advice and encouragement. Acknowledgment is also made of the use of the ROCKIT program, courtesy of Dr. Charles Metz of the University of Chicago.

REFERENCES

Brown, B. G., and A. H. Murphy, 1987: Quantification of uncertainty in fire-weather forecasts: Some results of operational and experimental forecasting programs. *Wea. Forecasting,* **2,** 190–205.

Green, D. M., and J. A. Swets, 1974: *Signal Detection Theory and Psychophysics.* Robert E. Kreiger Publishing, 479 pp.

Hamill, T. M., and D. S. Wilks, 1995: A probabilistic forecast contest and the difficulty in assessing short-range forecast uncertainty. *Wea. Forecasting,* **10,** 620–631.

Harvey, L. O., Jr., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.,* **120,** 863–883.

Katz, R. W., and A. H. Murphy, Eds.,1997: *Economic Value of Weather and Climate Forecasts.* Cambridge University Press, 222 pp.

Levi, K., 1985: A signal detection framework for the evaluation of probabilistic forecasts. *Organ. Behav. Hum. Decis. Processes,* **36,** 143–166.

Mason, I. M., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

——, 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteor. Mag.,* **37,** 75–81.

McMillan, N. A., and C. D. Creelman, 1991: *Detection Theory: A User's Guide.* Cambridge University Press, 407 pp.

Murphy, A. H., 1977: The value of climatological, categorical and probabilistic forecasts in the cost–loss ratio situation. *Mon. Wea. Rev.,* **105,** 803–816.

——, 1985: Decision making and the value of forecasts in a generalized model of the cost–loss ratio situation. *Mon. Wea. Rev.,* **113,** 362–369.

——, and R. L. Winkler, 1976: Reliability of subjective probability forecasts of precipitation and temperature. *Appl. Stat.,* **26,** 41–47.

——, and H. Daan, 1984: Impacts of feedback and experience on the quality of subjective probability forecasts: Comparison of results from the first and second years of the Zierikzee experiment. *Mon. Wea. Rev.,* **112,** 413–423.

——, and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting,* **7,** 435–455.

——, B. G. Brown, and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting,* **4,** 485–501.

Richardson, D. S., 2000: Skill and economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **126,** 649–668.

Stephenson, D. B., 2000: Use of the ''odds ratio'' for diagnosing forecast skill. *Wea. Forecasting,* **15,** 221–232.

Swets, J. A., 1986: Indices of discrimination or diagnostic accuracy. *Psychol. Bull.,* **99,** 100–117.

——, 1996: *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics, Collected Papers.* Lawrence Erlbaum Associates, 308 pp.

Wilks, D. S., 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.,* **8,** 209–219.

Winkler, R. L., and A. H. Murphy, 1979: The use of probabilities in forecasts of maximum and minimum temperatures. *Meteor. Mag.,* **108,** 317–329.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Milne, 2002: The economic value of ensemble weather forecasts. *Bull. Amer. Meteor. Soc.,* **83,** 73–83.